Journal of Epidemiology and Biostatistics (2000) Vol. 5, No. 5, 267–275

Probability of *Helicobacter pylori* infection based on IgG levels and other covariates using a mixture model

RM PFEIFFER, MH GAIL and LM BROWN
National Cancer Institute, Division of Cancer Epidemiology and Genetics, 6120 Executive Blvd, Bethesda, MD 20892-7244, USA

Background To use IgG antibody measurements to detect infection with *Helicobacter pylori (H. pylori),* one typically defines a cut-off value based on samples of persons presumed to be infected or uninfected. When there are no good 'gold standard' tests to determine infection status, or when laboratory conditions vary, it is useful to have a method based on the IgG measurements themselves to determine infection status.

Methods We present a two component mixture model to analyse serologic data on *H. pylori* infection. The mixing proportions correspond to the probability that a latent variable, the true, unknown infection status I of a person, is either 0 (uninfected) or 1 (infected). By using a logistic model for these probabilities, we are able to incorporate covariate information.

Results The model is applied to IgG data from Shandong, China. The distribution of the true infection status given the IgG value and a set of covariates is calculated using the IgG distribution function. An optimal cut-off point is found by minimising the probability of misclassification for the Shandong data. The optimal cut-off point is slightly lower than the pre-defined one.

Conclusions We contrast results from the mixture model with results from tabulations and from standard logistic regression that are based on fixed cut-points. The mixture model yields information on the probability that a person is truly infected as a function of IgG levels and covariates. In our data, the mixture model indicates that a slightly lower cut-off value than the pre-defined cut-point 1.0 can reduce misclassification rates.

Keywords *Helicobacter pylori*, mixture model, logistic regression, sensitivity, specificity.

Introduction

We are interested in quantifying the effects of factors that influence the prevalence of *H. pylori* infection in Linqu County, Shandong Province, China. The initial study population consisted of a nearly exhaustive census of 3411 subjects aged 35–69 from 13 randomly selected villages in Linqu. For each subject, *H. pylori* IgG antibody concentrations were measured using an enzyme-linked immunoassay (ELISA) procedure in 1994. In addition to the optical density measurements of IgG, covariate information on each person was obtained by questionnaire. Our analysis is based on the 3101 subjects with IgG measurements and complete questionnaire data. We use IgG to denote either the antibody class or the actual antibody optical-density measurements.

Standard statistical approaches to investigating factors that affect the prevalence of *H. pylori* infection, such as contingency table analyses and logistic regression, employ an operational definition of 'infected', namely that the IgG optical density exceed a given cut-off value. Based on IgG data from samples of verified infectees and from very probably uninfected children, the cut-off point $IgG \geq 1.0$ has been used for the classification 'infected'[1]. However, the measurement of optical density is sensitive to slight changes in the laboratory procedure. Thus relying on a fixed cut-off value may affect the misclassification rate. For example, if the proper threshold value were 1.2 instead of 1.0, the derived prevalence would decrease from 65.8% to 63.9% in our data.

In this paper we analyse the data by fitting a two-component mixture model to a properly chosen transformation of the IgG measurements. The probability of being in one state or the other, namely the true, unobservable infection status, is modelled by a logistic regression that allows us to incorporate the covariate information. Our approach resembles that of Thompson *et al.*[2], who applied mixture models to the diagnosis of diabetes based on plasma glucose level. We think that the application to *H. pylori* presented here is interesting in its own right and that it illustrates a technique that may be useful to epidemiologists.

Correspondence to: RM Pfeiffer, Division of Epidemiology and Genetics, 6120 Executive Blvd, EPS/8017, Bethesda, MD, 20892-7244, USA.

Received 10 January 2000
Revised 22 March 2000
Accepted 12 June 2000

The mixture model approach has several potential advantages. First, we need not rely on an external definition of a cut-off value to classify each observation. Second, the continuous nature of the IgG data is used to its full extent and we obtain a complete description of the distribution of the IgG values in the presence of the covariates. This enables us to calculate $P(\text{infected}|IgG, z)$, the probability of being truly infected given the IgG reading and a set of covariates z.

We define the logistic mixture model formally and briefly review the methods of inference for the model. We then present the results of applying the mixture model to the Shandong data. We compare those results with results obtained from contingency tables and logistic regression, $P(X = 1|z) = \exp(z'\beta)/[1 + \exp(z'\beta)]$, based on observations of $X = I(IgG \geq 1.0)$, where I, the indicator function, is one if the argument is true and zero otherwise. In the last section we comment on the strengths and weaknesses of the mixture model method.

Model formulation and estimation

Mixture models (see References 3 and 4 for good introductions) were developed as a way of analysing data that arise from two or more distinct data-generation processes. The two component logistic mixture model we consider in this paper was motivated by the histogram of transformed IgG values, $\ln(IgG + 0.5)$, shown in Figure 1. This histogram suggests a mixture of two densities, the one to the right corresponding to the IgG values of infected individuals and the one to the left corresponding to IgG values of uninfected subjects. We characterise the logistic mixture model as follows. The data consist of the pairs (Y_j, z_j) for $j = 1, \ldots, n$, where Y_j denotes the observed IgG measurement and z_j a $p \times 1$ covariate vector for the j-th person in the study. Each person is in one of two latent true infection states, which we label as state $I_j = 1$ ('infected') and state $I_j = 0$ ('uninfected'). The state probabilities for the j-th observation depend on z_j through a logistic regression:

$$P[I_j = 1 \mid z_j] = p(z_j; \beta) = \frac{\exp(z_j'\beta)}{1 + \exp(z_j'\beta)}.$$

The first component of z_j is unity and corresponds to an intercept. Given z, the probability density function of Y is given by the mixture model:

$$g(y|z, \theta) = f(y; \alpha_0) \cdot (1 - p(z;\beta)) + f(y; \alpha_1) \cdot p(z; \beta) \quad (1)$$

where $f(.,\alpha)$ is a parametric density function, such as the normal density and $\theta = (\alpha_0, \alpha_1, \beta)$. We interpret $f(.,\alpha_0)$ to be the density of the IgG values (or a known transformation of the IgG values) that corresponds to persons in the 'uninfected' state and $f(.,\alpha_1)$ to be the density of the IgG values that corresponds to subjects in the 'infected' state.

A more general model would allow $f(.,\alpha_0)$ and $f(.,\alpha_1)$ to depend on covariates[2]. Among the variables we consider (see below, Model selection and comparison of the mixture and logistic models), the only covariate that might plausibly influence antibody concentration conditional on latent infection status is age. Using the extended model[2] and allowing for separate age effects among infected and uninfected subgroups, we regressed the means of the mixture densities on age. The age effects were small and not statistically significant, we therefore assume that the distribution of the IgG level for given infection status does not depend on any covariates, as in equation (1).

The density of the $\ln(IgG, + 0.5)$ values corresponding to the 'infected' state is the component density with the larger mean in Figure 1. The density component with the smaller mean corresponds to $\ln(IgG + 0.5)$ for the 'uninfected' state. The symmetric densities in the histogram of the transformed data $y_i = \ln(IgG_i + 0.5)$ are suggestive of a mixture of two normal densities. We chose this transformation to render the component densities approximately normal. Formal tests (see below) suggest that the mixed normal model based on this transformation fits the data adequately. Although the component densities are affected by transformations of y, the mixing proportion $p(z;\beta)$ is not.

Thus, the full parametric model we use is given by (1) where $y = \ln(IgG + 0.5)$ and f denotes the normal density with mean μ_k and variance ϕ_k^2, for infection status $k = 0$ or 1. In the next section we apply the model stated in (1) to the *H. pylori* data-set. For comparison, we also study the case of no covariates, namely a standard mixture

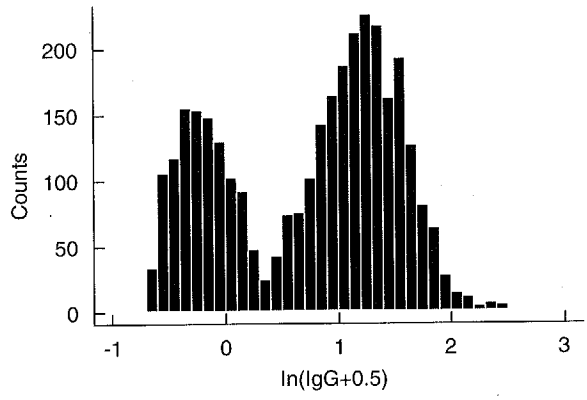

***Fig. 1** Histogram of ln(IgG + 0.5).*

model with constant state probabilities, $P[I_j = 1] = p$, or, using the logistic parameterisation,

$$P[I_j = 1] = \exp(\beta_1) / [1 + \exp(\beta)].$$

We use the EM algorithm to find maximum likelihood estimates for α_0, α_1 and β. The EM algorithm is an iterative procedure to find maximum likelihood estimates in problems with missing or unobserved data[5]. In this instance the unobserved quantity is the latent infection status, or its indicator I_j. The algorithm is used to maximise the log likelihood given by

$$\sum_{j=1}^{n} \log g(y_j|z_j; \theta).$$

The implementation and interpretation of the EM algorithm in the context of mixture models has been discussed by several authors[3,4,6]. The derivation of the algorithm for logistic mixtures such as equation (1) is given in detail elsewhere[7,2]. We note that the maximisation step in the EM procedure is often easy to perform. In particular, if the component densities are from an exponential family (e.g. normal, Poisson, beta, binomial), many statistical software packages will perform the M-step through a generalised linear model regression routine that accepts the E-step estimates $1 - p(z; \beta)$ and $p(z; \beta)$ as weights. In our model the component densities are normal and do not depend on covariates; we therefore obtained closed-form solutions for the mean and variance estimates of the component densities in the M-step, which makes EM particularly simple.

If (Y_j, Z_j) are independent and identically distributed and under mild conditions on $f(.,\alpha)$, the maximum likelihood estimate, $\hat{\theta}_n$, has an asymptotic normal distribution and satisfies:

$$\sqrt{n}\,(\hat{\theta}_n - \theta^*) \rightarrow N(\underline{0}, Q^{-1})$$

where $\theta^* = (\alpha_0^*, \alpha_1^*, \beta^*)$ denotes the true parameters and

$$Q = E\left[\frac{\partial \log g(y|z; \theta)}{\partial \theta} \frac{\delta \log g(y|z; \theta)}{\partial \theta} \bigg|_{\theta^*} \right].$$

To get the needed variances, we approximated the information matrix Q by

$$\hat{Q}_n = \frac{1}{n} \sum_{j=1}^{n} \left[\frac{\partial \log g(y_j | z_j; \theta)}{\partial \theta} \frac{\partial \log g(y_j | z_j; \theta)}{\partial \theta} \bigg|_{\hat{\theta}_n} \right].$$

Alternatively, Q can be approximated by

$$\tilde{Q}_n = \frac{1}{n} \sum_{j=1}^{n} H_j,$$

where H_j denotes the negative Hessian of $\log g(Y_j|z_j;\theta)$ obtained through numerical differentiation at $\hat{\theta}$. In smaller samples $\tilde{Q}_n$ may be preferable, but in our example $\hat{Q}_n$ and $\tilde{Q}_n$ gave nearly identical results.

Results

Model selection and comparison of the mixture and logistic models

The covariates $z = (z_1, .., z_6)$ we considered are: $z_1 = 1$, an intercept term; z_2 = 'hands washed', in four ordered categories: always (0), more than half the time (1), less than half the time (2), never (3); z_3 = 'age', in three categories: ≤ 44 years (0), 45–54 years (1), ≥ 55 years (2); z_4 = 'total number of children in the household', in three categories, 0, 1, 2 (more than one child); z_5 = 'water source', in three ordered categories, deep private well (1), shallow private well or deep village well (2), other sources (3); and z_6 = 'smoking', no (0) , yes (1).

Of the 3411 subjects in the study 3101 had complete data on IgG measurements and all the covariates. Fifty-four subjects had missing IgG values, 213 had completely-missing covariate information and 133 had partially-missing covariate information. We examined subjects with missing data to determine if they were similar to those with complete data, conditional on the covariates included in the model. For those with only IgG and age data, the IgG distribution was similar within age group to the IgG distribution for the subjects with complete data. Similar results were obtained for those with IgG and 'total number of children in the household' and for those with completely missing covariate information. Those with missing values seemed to be similar to those with complete data within categories included in later models, we therefore felt justified in confining our analysis to the 3101 subjects with complete data.

In Table 1 we present point estimates and standard errors (in parentheses) for several models. The columns headed LM1 and LM2 correspond respectively to results for logistic mixtures with the full covariate vector and with covariates z_2 and z_6 omitted. The CM column denotes the mixture with constant state probabilities (i.e. no covariates except Z_1). 'logit 1' and 'logit 2' correspond respectively to logistic regression of $X = I(IgG \geq 1.0)$ on a full covariate vector and on the vector with z_2 and z_6 omitted.

The saturated model with six logistic parameters had log-likelihood −2863.9 which only decreased to −2865.4 when the two covariates with non-significant Wald tests, z_2 and z_6, were eliminated (see LM2 in Table 1). This change in log-likelihood corresponds to a non-significant χ^2 value of 3.0 on two degrees of freedom ($p = 0.22$). Thus, we use the logistic model LM2 in

Table 1 Three mixture models for ln ($IgG + 0.6$) and two logistic models for the event $IgG \geq 1.0$

Parameter	CM[a]	LM1[b]	LM2[b]	logit 1[c]	logit 2[c]
μ_1	1.2141(0.0096)	1.2132 (0.0096)	1.2131 (0.0096)	NA[d]	NA
ϕ_1	0.3972 (0.0072)	0.3981 (0.0073)	0.3982 (0.0073)	NA	NA
μ_0	−0.2061 (0.0083)	−0.2062 (0.0083)	−0.2069 (0.0083)	NA	NA
ϕ_0	0.2423 (0.0080)	0.2417 (0.0080)	0.2417 (0.0080)	NA	NA
β_1	0.7042 (0.0403)	0.7167 (0.0408)	0.7152 (0.0407)	0.6631 (0.0381)	0.6615 (0.0380)
β_2	NA	0.0476 (0.0416)	NA	0.0426 (0.0387)	NA
β_3	NA	−0.0805 (0.0408)	−0.0815 (0.0408)	−0.0755 (0.0391)	−0.0764 (0.0391)
β_4	NA	0.0793 (0.0415)	0.0785 (0.0414)	0.0706 (0.0389)	0.0699 (0.0389)
β_5	NA	0.1773 (0.0401)	0.1767 (0.0401)	0.1614 (0.0388)	0.1605 (0.0388)
β_6	NA	−0.0489 (0.0395)	NA	−0.0502 (0.0381)	NA
Log likelihood	−2879.9	−2863.9	−2865.4	−1977.1	−1978.5

[a] CM denotes the mixture with constant state probabilities (i.e. no covariates except z_1).

[b] Columns LM1 and LM2 correspond respectively to results for logistic mixtures with the full covariate vector and with covariates z_2 and z_6 omitted.

[c] 'logit 1' and 'logit 2' correspond respectively to logistic regression of $X = I(IgG \geq 1.0)$ on a full covariate vector and on the vector with z_2 and z_6 omitted.

[d] NA = not available.

subsequent analyses. The model CM has log likelihood −2879.9, which corresponds to a χ^2 of 29.0 on three degrees of freedom ($p = 2 \cdot 10^{-6}$). Thus there is evidence that the probability of being infected depends on z_3, z_4 and z_5.

The standard logistic model, logit 1, indicates that the variables z_1, z_3, z_4 and z_5 are needed to predict $X = I(IgG \geq 1.0)$, but that z_2 and z_6 are not needed (see logit 2, Table 1). Hence, the analyses based on the mixture and the logistic models lead to the same selection of covariates for predicting infection status. These findings are consistent with the literature[8]. The number of children in the household is a measure of over-crowding that has been consistently related to *H. pylori* infection. Subjects who obtain their water from wells that are more apt to be contaminated with material containing *H. pylori* have been found to be at higher risk of infection. Age has a slightly protective effect in this population, as some previously infected individuals may suffer atrophic change in gastric mucosa, creating an inhospitable environment for *H. pylori*[9].

The coefficients in the logistic regression model $P(X = 1|z) = \exp(z'\beta)/[1 + \exp(z'\beta)]$ have a different interpretation than the parameters β in the mixture model. In the mixture model the logistic probability corresponds to the probability of the true, unobservable infection status, I, whereas in the standard logistic regression the probability of the observable event $X = I(IgG \geq 1.0)$ is estimated. If the operational definition of infection, $IgG \geq 1.0$, coincided perfectly with the 'true' latent state $I = 1$, then the mixture and logistic models would yield identical prediction of infection status, and the estimates of β_i would be equal, apart from random variation. In fact, the estimates for LM2 and logit 2 are not drastically different, but they are not identical. For example, LM2 indicates that a change from water source 1 to water source 2 is associated with an increase in the odds of $I = 1$ of $\exp(0.1767) = 1.19$, whereas logit 2 indicates that the corresponding increase in the odds of $IgG \geq 1.0$ is $\exp(0.1605) = 1.17$.

Both models give very similar prediction of $P(IgG \geq 1.0|z)$ (see asterisks in Figure 2). Whereas such predictions based on the logit 2 model are immediate, to compute $P(IgG \geq 1.0)$ from LM2, one needs to integrate equation (1). The crosses in Figure 2 depict estimates of $P(I = 1|z)$ from LM2 plotted against estimates of $P(IgG \geq 1.0|z)$ from logit 2 for various choices of z. This plot indicates that the former tends to exceed the latter. If the assumptions underlying the logistic mixture model are correct, this analysis suggests that the operational definition, $IgG \geq 1.0$, slightly under-estimates the proportion infected. The unconditional estimate of $P(I = 1) = 0.6698$ from LM2 is obtained as $\sum_z P(I = 1 \mid z)P(z)$, where $P(z)$ is the observed proportion of the population at covariate level z. The proportion truly infected is estimated from CM as 0.6690, in very close agreement. Both these estimates of $P(I = 1)$ exceed the observed proportion with $IgG \geq 1.0$, namely $0.6582 = 2041/3101$.

These calculations suggest that the cut-off value of $IgG \geq 1.0$ is slightly too high.

Similar conclusions can be drawn from Table 2. The entries in Table 2 are $P(IgG > 1.0|z)$ from LM2 and logit 2 respectively and $P(I = 1|z)$ from LM2 (in parenthesis) for all possible combinations of the covariates. Note that $P(IgG \geq 1.0|z)$ estimated from LM2 is very close to the estimate obtained from logit 2. In contrast, the estimate of $P(I = 1|z)$ from LM2 consistently exceeds the estimates of $P(IgG \geq 1.0|z)$. It is also apparent from Table 2 that water source (z_5) has a greater impact on the probability of infection, judged by either $P(IgG \geq 1.0|z)$ or by $P(I = 1|z)$ than does age (z_3), or number of children in the household (z_4).

We find that cut-off value c^* for IgG that minimises the probability of misclassification if the estimated latent mixture model is true. To determine c^* we minimise the probability of misclassification:

$$[1 - P(I=1)] \int_c^{\infty} f(y;\alpha_0)dy + P(I=1) \int_{-\infty}^{c} f(y;\alpha_1)dy, \quad (2)$$

where the α_i, $i = 0,1$ and $P(I = 1)$ are replaced by their estimates. The optimal cut-off point for the above criterion under the mixture model is $c^* = 0.87$ (after transforming back to the original IgG scale), with a probability of misclassification of 0.013; the cut-off point of $c = 1.0$ (that corresponds to a cut-off point of 0.405 on the transformed scale) yields a misclassification probability of 0.016. Using the new cut-off point, we classify 2061 of the 3101 subjects as infected; the corresponding number based on $IgG \geq 1.0$ is 2041. Though the change in the cut-off value may seem substantial, only 20/3101 = 0.65% of the population are classified differently. To obtain the variance of the cut-off point we performed a bootstrap with 500 repetitions by resampling the (y_j, z_j) data. The mean value for the bootstrap repetitions is 0.8796 and the 95% confidence interval (CI) based on the bootstrap sample standard deviation is (0.740, 1.034). Thus, the pre-defined cut-off point of 1.0 falls just within the CI.

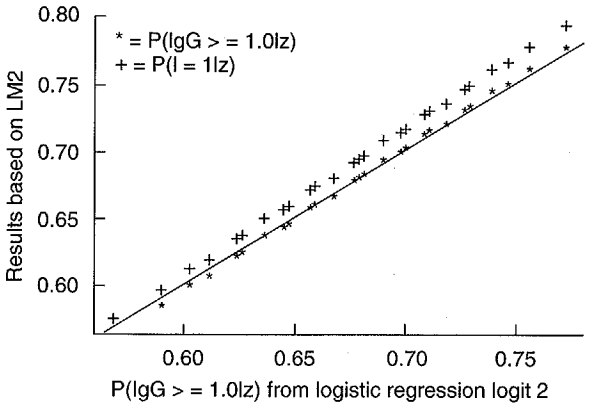

Fig. 2 *Plots of estimates of $P(IgG \geq 1.0|z)$ and $P(I = 1|z)$ from the mixture model LM2 against the estimate of $P(IgG \geq 1.0|z)$ from the standard logistic model, logit2, for various choices of covariates. The solid line denotes the equiangular line, ordinate = abscissa.*

To maximise the sum of specificity and sensitivity, we find the value c^* that maximises

$$\int_{-\infty}^{c} f(y;\alpha_0)dy + \int_c^{\infty} f(y;\alpha_1)dy.$$

This yields a cut-off value of 0.95 and a classification of 2047/3101 individuals as infected. Thus both optimality criteria result in cut-off values for IgG < 1.0. The corresponding bootstrap CI is (0.846–1.122) which also includes the pre-defined cut-off point.

The logistic mixture model gives a more complete description of the IgG data than the standard logistic model, which only predicts the proportion with $IgG \geq 1.0$. For example, for $z = (z_1, z_3, z_4, z_5)$ (1, 2, 0, 2), the estimated density of IgG agrees well with the observed histogram of $\ln(IgG + 0.5)$ values (Figure 3). It is also

Table 2 Estimates of $P(IgG \geq 1.0 \mid z)$ from LM2 and logit 2 respectively, followed by an estimate of $P(I = 1 \mid z)$ from LM2 (in parenthesis)

z_5	z_3 / z_4	0	1	2
1	0	0.608, 0.612 (0.619)	0.644, 0.645 (0.656)	0.678, 0.677 (0.691)
	1	0.585, 0.590 (0.596)	0.622, 0.624 (0.634)	0.658, 0.657 (0.670)
	2	0.563, 0.568 (0.572)	0.600, 0.603 (0.611)	0.637, 0.637 (0.648)
2	0	0.667, 0.668 (0.679)	0.700, 0.699 (0.713)	0.731, 0.728 (0.745)
	1	0.646, 0.648 (0.658)	0.680, 0.679 (0.693)	0.712, 0.710 (0.726)
	2	0.625, 0.627 (0.636)	0.660, 0.659 (0.672)	0.693, 0.690 (0.707)
3	0	0.721, 0.719 (0.735)	0.750, 0.747 (0.765)	0.777, 0.773 (0.792)
	1	0.702, 0.700 (0.716)	0.733, 0.729 (0.747)	0.761, 0.757 (0.776)
	2	0.683, 0.681 (0.696)	0.715, 0.712 (0.728)	0.744, 0.740 (0.759)

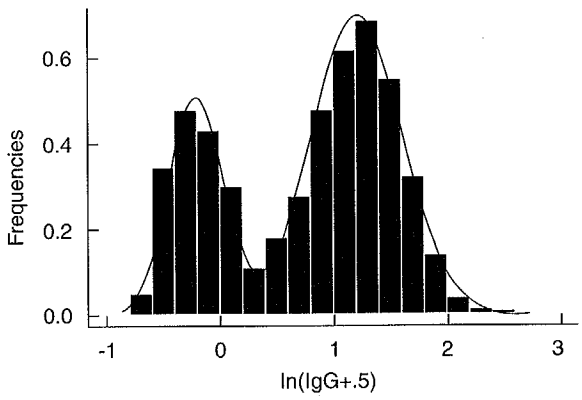

Fig. 3 Fitted density from the mixture model LM2 and observed histogram for z = (1, 2, 0, 2).

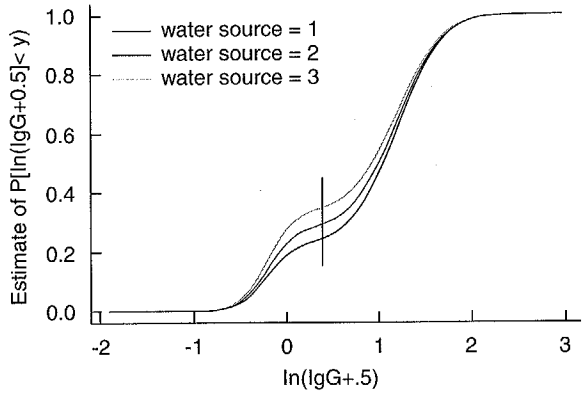

Fig. 4 Estimates of the cumulative distribution function of ln(IgG + 0.5) from the mixture model LM2 for water source z_5 = 1, 2 and 3. The vertical line corresponds to the cut-off IgG = 1.0.

possible to predict the density and distribution of IgG for any subset of fixed covariates. To study water source (z_5), for example, we calculate:

$$g(y|z_5) = \sum_{z_3,z_4} g(y|z_3, z_4, z_5)\, P(z_3, z_4|z_5)$$

where $g(y|z) = [1-p(z;\beta)]\, f(x;\mu_0,\phi_0) + p(z;\beta)\, f(x;\mu_1, \phi_1)$, and $P(z_3, z_4|z_5)$ is estimated from the corresponding sample frequency. The corresponding probability distribution functions are obtained by integrating $g(y|z_5)$. The distribution functions for $\ln(IgG + 0.5)$ are shown for water sources 1,2, and 3 in Figure 4. In particular, the distribution functions give the probabilities $P(IgG \geq 1.0|\text{water source}) = P(\ln(IgG + 0.5) \geq 0.4055 |\text{water source})$, which are indicated by a vertical bar in Figure 4. The ordinal values in Figure 4 yield the proportions $P[\ln(IgG + 0.5) \geq 0.4055|\text{water source} = 1] = 0.6518$, $P[\ln(IgG + 0.5) \geq 0.4055|\text{water source} = 2] = 0.7068$, and $P[\ln(IgG + 0.5) \geq 0.4055|\text{water source} = 3] = 0.7559$.

A nice feature of the logistic mixture model is its ability to calculate the probability of $I_j = 1$ given z and $y_j = \ln(IgG_j + 0.5)$. Indeed, from (1) we get

$$P(I_j=1|y_j,z_j)=\frac{p(z_j;\beta)f(y_j;\mu_1,\phi_1)}{(1-p(z_j;\beta))f(y_j;\mu_0,\phi_0)+p(z_j;\beta)f(y_j;\mu_1,\phi_1)}.$$

For $z = (1,0,0,1)$ and $z = (1,0,0,3)$, $P(I_j = 1|y_j, z_j)$ rises rapidly in the region $0.2 < \ln(IgG + 0.5)$ (see Figure 5). The optimal cut-off for minimising equation (2) when discriminating $I_j = 1$ from $I_j = 0$, is to select $I_j = 1$ if $P(I_j = 1|y_j, z_j) \geq 0.5$, and $I_j = 0$ otherwise. From Figure 5, the optimal cut-off values are $\ln(IgG + 0.5) \geq 0.3$ and 0.35 respectively, for $z_5 = 3$ and 1. These values correspond to IgG values of 0.85 and 0.92 respectively, rather than 1.0.

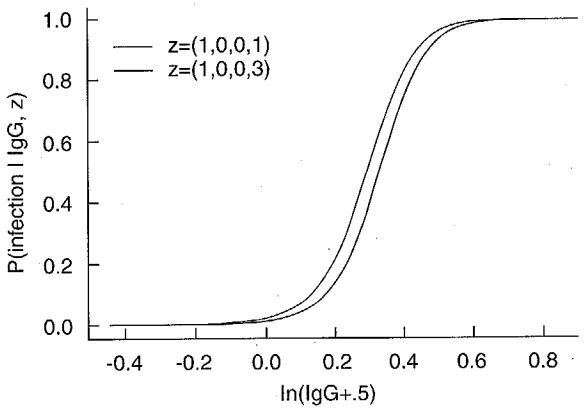

Fig. 5 Estimates of P(I = 1|z, IgG) from the logistic mixture model LM2 for z = (1, 0, 0, 1) and for z = (1, 0, 0, 3).

Although these differences are not dramatic, they illustrate the fact that one might wish to use different cut-off values to predict an individual's true infection status, depending on his or her covariate values.

Goodness-of-fit

To test the goodness-of-fit of the logistic mixture model, we compared the observed proportions with $IgG \geq 1.0$ with predicted proportions (Table 3).

The upper number in each cell of Table 3 is the observed number of subjects with $IgG \geq 1.0$ (out of the total number of subjects in the cell). The numbers in the next two rows are expected counts of $IgG \geq 1.0$ based on LM2 and logit 2 respectively, and the number in the last row of each cell is the estimated count with $I = 1$, based on LM2.

We used as the goodness-of-fit criterion the sum over the 21 non-empty cells in Table 3 of $n(\hat{p}_0-\hat{p})^2/\hat{p}(1-\hat{p})$

Table 3 Observed and expected quantities for testing goodness-of-fit

	$z_5 = 1$				$z_5 = 2$				$z_5 = 3$			
z_3 / z_4	**0**	**1**	**2**	**Row total**	**0**	**1**	**2**	**Row total**	**0**	**1**	**2**	**Row total**
0	86 (136)	50 (78)	6 (8)	142	0 (0)	0 (0)	0 (0)	0	286 (396)	83 (111)	8 (11)	377
LM2	82.6369	50.2008	5.4227		0	0	0		285.3489	83.2243	8.5425	
logit 2	83.2314	50.3290	5.4172		0	0	0		284.9114	82.9519	8.5059	
$I = 1$	84.1334	51.1366	5.5262		0	0	0		290.9424	84.8817	8.7149	
1	56 (95)	11 (23)	0(2)	67	0 (0)	0 (0)	0 (0)	0 (0)	151 (214)	13 (23)	1 (1)	165
	55.6054	14.3101	1.3150		0	0	0		150.2153	16.8478	0.7608	
	56.0874	14.3602	1.3143		0 (0)	0 (0)	0 (0)		149.9993	16.7882	0.7571	
	56.5922	14.5724	1.3390		0	0	0		153.1278	17.1803	0.7760	
2	29 (60)	6 (8)	1 (2)	36	754 (1185)	233 (362)	43 (53)	1030	199 (300)	24 (32)	1(1)	224
	33.7561	4.8016	1.2730		740.3044	238.8929	36.7415		204.7546	22.8614	0.7441	
	34.1063	4.8240	1.2731		742.7687	238.7368	36.6086		204.5117	22.7774	0.7401	
	34.3416	4.8880	1.2966		753.9036	243.3993	37.4500		208.6765	23.3079	0.7589	

where $\hat{p}_o$ is the observed proportion with $IgG \geq 1.0$, n is the number of observations, and $\hat{p}$ is the proportion estimated to have $IgG \geq 1.0$ in a cell under the mixture model, LM2. This criterion is the sum over the cells of the Pearson χ^2 for the two binomial outcomes in each cell and has degrees of freedom equal to the number of non-empty cells minus the number of fitted parameters. For LM2, the Pearson χ^2 is 18.2609 on $21 - 8 = 13$ degrees of freedom, which indicates adequate fit ($p = 0.15$). The fit can also be assessed informally by comparing observed and predicted distributions of $\ln(IgG + 0.5)$ for various values of z. For example, for $z = (1, 2, 0, 2)$, the predicted density seems to fit the observed histogram well (see Figure 3). Formal procedures are available to assess the fit of a distribution function to the empirical distribution function when the parameters are estimated[10], but this approach is not pursued here.

Data in Table 3 with $\hat{p}$ predicted from the logistic model yield a χ^2 of 18.1885 with $21 - 4 = 17$ degrees of freedom ($p = 0.38$). Thus the logistic model fits well. A more refined examination of the distribution of IgG is not possible for the logistic model, which only predicts $P(IgG \geq 1.0|z)$.

Discussion

In this paper we use a two component mixture model to analyse serologic data on *H. pylori* infection. The stated probabilities correspond to the probability that a latent variable, the true, unknown infection status I of a person, is either 0 (uninfected) or 1 (infected). By using a logistic model for those probabilities, we are able to incorporate covariate information. The results obtained from the mixture model are compared with the results obtained from tables and from a standard logistic regression applied to the data $X = I(IgG \geq 1.0)$.

The standard logistic model and the mixture model fit the data very well, agree in the selection of the important covariates z_3, z_4 and z_5, and agree in predicting $P(IgG \geq 1.0|z)$. However, estimates of $P(I_j = 1.0|z)$ tend to exceed $P(IgG \geq 1.0|z)$, suggesting that a slightly lower cut-off value would be appropriate for these data. Using plots like Figure 2, or tabulations, such as Table 2, it is possible to search for covariate combinations with huge discrepancies between $P(I = 1.0|z)$ and $P(X = 1.0|z)$; such an examination might suggest covariate patterns for which the usual cut-off value is misleading. In the present example, the discrepancies were quite regular though somewhat larger for persons using public surface water sources ($z_5 = 3$) and households with more than one child ($z_4 = 2$). These discrepancies could be explained largely by a simple calibration problem that suggests the need for a slightly lower cut-off point.

The mixture model analysis suggests, however, that the prespecified cut-off value 1.0 works quite well in these data. In particular, prevalence estimates are affected very little by lowering the cut-off point to 0.87, the estimated optimal cut-off value from the mixture model LM2, and the confidence interval on the estimated optimal cut-off, (0.74,1.03), includes the pre-specified cut-off 1.0.

Thompson *et al.*[2] use a similar mixture model for glucose levels to diagnose diabetes in the presence of covariates. Their paper also illustrates the use of the mixture method to find an appropriate cut-off point when there is no 'gold standard' of disease status. Thompson *et al.* allowed for the incorporation of covariates in the component densities in the prediction of latent disease status; their model is thus slightly more general than the model presented in this paper. For example, Thompson *et al.* found obesity to be a good predictor of glucose levels among those thought to be diabetic, but not among those thought not to be diabetic. Although this generalisation can yield valuable insights and, potentially, improve model fit, it was unnecessary for our problem, because the covariates considered were thought to influence the chance of being infected, but not the IgG distributions conditional on infection status. When we used age as a predictor of density means in the more general model, the corresponding estimated age effects were small and not statistically significant. Thus, the more general model was not needed in our case.

The main advantage of the mixture approach is that an *a priori* classification of the observations into an infected and uninfected group based on a cut-off point is unnecessary. A complete set of covariate-specific probability distributions of the IgG values is also obtained. Using the IgG distribution function, $P(I_j = 1.0|y_j z_j)$ can be calculated and provides additional insight into the relationship between the unobservable infection status and the observable IgG measurement of a person in the presence of the covariates. Moreover, it may be more realistic to estimate the probability of infection as a function of IgG, rather than assert that the subject is infected or not based on a single cut-point.

A weakness of this latent mixture model is that it depends on parametric assumptions and that the estimation procedures are more complex than standard logistic regression, as the parameters of the component densities have to be estimated. We have indicated some ways to test goodness-of-fit, and in our example the fit appears to be quite good. Further work would be useful to test sensitivity of the results to departure from the parametric assumptions and to extend the model to mixtures with nonparametric densities. The MATLAB 5.0 program used to analyse these data is available on request from the first author.

Acknowledgements

We thank Wei-Cheng You for providing the data and Barry Graubard and Neal Jefferies for helpful suggestions and remarks.

References

1 Zhang L, Blot WJ, You W *et al. Helicobacter pylori* antibodies in relation to precancerous gastric lesions in a high-risk Chinese population. *Cancer Epidemiol Biomark Prevent* 1996;5:627–30.

2 Thompson TJ, Smith PJ, Boyle JP. Finite mixture models with concomitant information: assessing diagnostic criteria for diabetes. *Appl Stat* 1998;47:393–404.

3 McLachlan GJ, Basford KE. *Mixture models: inference and applications to clustering.* New York: Marcel Dekker, Inc., 1988.

4 Titterington DM, Smith AFM, Makov UE. *Statistical analysis of finite mixture distributions.* New York: Wiley, 1985.

5 Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm [with discussion]. *J R Statist Soc B* 1977;39:1–39.

6 McLachlan GJ, Krishnan T. *The EM algorithm and extensions.* New York: Wiley, 1997.

7 Jeffries N. Logistic mixtures of generalized linear model time series. [Ph.D. Thesis] Maryland:University of Maryland, 1998.

8 Brown LM. *Helicobacter pylori*: epidemiology and routes of transmission. *Epidemiol Rev.* In press.

9 Forman D. *Helicobacter pylori* infection and cancer. *Br Med Bull* 1998;54:71–8.

10 Csörgö M, Révész P. *Strong approximations in probability and statistics.* New York: Academic Press, Inc., 1981.